

# Statistical Potential-Based Amino Acid Similarity Matrices for Aligning Distantly Related Protein Sequences

Yen Hock Tan,<sup>1</sup> He Huang,<sup>2</sup> and Daisuke Kihara<sup>1,2,3,4\*</sup>

<sup>1</sup>Department of Computer Sciences, College of Science, Purdue University, West Lafayette, Indiana

<sup>2</sup>Department of Biological Sciences, College of Science, Purdue University, West Lafayette, Indiana

<sup>3</sup>Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, Indiana

<sup>4</sup>The Bindley Bioscience Center, College of Science, Purdue University, West Lafayette, Indiana

**ABSTRACT** Aligning distantly related protein sequences is a long-standing problem in bioinformatics, and a key for successful protein structure prediction. Its importance is increasing recently in the context of structural genomics projects because more and more experimentally solved structures are available as templates for protein structure modeling. Toward this end, recent structure prediction methods employ profile–profile alignments, and various ways of aligning two profiles have been developed. More fundamentally, a better amino acid similarity matrix can improve a profile itself; thereby resulting in more accurate profile–profile alignments. Here we have developed novel amino acid similarity matrices from knowledge-based amino acid contact potentials. Contact potentials are used because the contact propensity to the other amino acids would be one of the most conserved features of each position of a protein structure. The derived amino acid similarity matrices are tested on benchmark alignments at three different levels, namely, the family, the superfamily, and the fold level. Compared to BLOSUM45 and the other existing matrices, the contact potential-based matrices perform comparably in the family level alignments, but clearly outperform in the fold level alignments. The contact potential-based matrices perform even better when suboptimal alignments are considered. Comparing the matrices themselves with each other revealed that the contact potential-based matrices are very different from BLOSUM45 and the other matrices, indicating that they are located in a different basin in the amino acid similarity matrix space. *Proteins* 2006;64:587–600. © 2006 Wiley-Liss, Inc.

**Key words:** amino acid similarity matrices; distantly related protein sequences

## INTRODUCTION

There are two types of approaches that are used to predict tertiary structure of proteins from their amino acid sequence. The first one is to simulate the folding process of a protein by applying a physics-based force field. This kind of force field models physical interactions between atoms in proteins and also between solvent molecules.<sup>1–3</sup> Another approach is to use the evolutionary information learned from the observation of the increasing number of

experimentally determined sequences and structures of proteins, applying the knowledge that homologous proteins have a similar fold.<sup>4,5</sup> This evolutionary principle is extended to state that protein local fragments of a similar sequence, which do not necessarily have an evolutionary relationship, have a similar local structure.<sup>6–8</sup> Of course, these two approaches are not necessarily mutually exclusive, and they are often combined as terms in the scoring function of “ab initio” or “de novo” protein structure prediction algorithms.<sup>9–12</sup> Although both approaches have been making steady progress in the past years, the latter approach is more practical at this point in the sense that it can constantly make good predictions for larger proteins when an appropriate template structure for the query sequence is available.<sup>13,14</sup> Two categories of protein structure prediction methods, namely, homology (comparative) modeling<sup>15</sup> and threading,<sup>16–20</sup> here referred to as template-based protein structure prediction methods, clearly fall into the latter approach.

For template-based protein structure prediction methods, a key for accurate prediction is the quality of sequence alignments used either as the initial step for the modeling procedure or as a dominant scoring term to construct an alignment between the query sequence and the template protein structure. In comparative modeling, a severe error in the initial alignment cannot be recovered by subsequent steps of side-chain and main-chain optimization.<sup>15</sup> Recent threading algorithms have improved their ability to recognize the correct fold from a template structure database, but alignment errors between the query sequence and the template structure are still a major problem.<sup>21,22</sup> Even for ab initio methods, which fold a simplified protein model from scratch, the alignment procedure is an essential component, because they often use local structure information obtained from the initial database search.<sup>9–12,23,24</sup>

A promising way to improve the quality of alignment between a query protein sequence and a template is to employ multiple sequence alignments to enrich evolutionary information.<sup>25–29</sup> Actually, some of the recent success-

\*Correspondence to: Daisuke Kihara, Department of Biological Sciences, College of Science, Lilly Hall, 915 West State Street, Purdue University, West Lafayette, IN 47907. E-mail: dkihara@purdue.edu

Received 9 November 2005; Revised 17 February 2006; Accepted 9 March 2006

Published online 23 June 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21020

ful threading algorithms essentially employ only multiple sequence alignment information in the form of profile–profile alignments, where a profile is created for the query protein and the template protein to be aligned.<sup>30–33</sup> It should be noted that the recent improvement of profile–profile alignments is also attributed to the rapid growth of sequence data available to construct profiles<sup>34</sup> and the use of PSI-BLAST,<sup>35</sup> which can collect more distant sequences in a database search.

More fundamentally, scoring matrices used for creating sequence alignments and profiles can be improved. A template-based protein structure prediction algorithm is aimed at aligning two proteins with a barely recognizable evolutionary relationship (i.e., superfamily level similarity) and even protein pairs of no evolutionary relationship but sharing the same fold (fold level similarity). Apparently, to achieve this, a prediction algorithm should use a scoring scheme that reflects structural similarity but does not solely rely on the homologous relationship of proteins. By definition, commonly used amino acid similarity matrices (AA matrices) that are derived from a set of homologous proteins, such as the BLOSUM<sup>36</sup> or PAM<sup>37</sup> series, do not work well for proteins of fold level similarity. Note here that if a better AA matrix for distantly related proteins is developed, it will benefit not only the construction of pairwise but also profile–profile alignments because the new AA matrix could make a better profile itself for the query and template protein sequences.

Several AA matrices have been proposed for aligning distantly related protein sequences. A typical method is to use structurally aligned protein sequences as the reference.<sup>38,39</sup> Another technique is to optimize an AA matrix numerically by using an optimization algorithm.<sup>40,41</sup>

In this study, we have derived a series of novel AA matrices that are calculated from the similarity of the contact propensity between pairs of amino acid residues. The contact propensity of amino acids can be described in the form of a pairwise residue contact potential.<sup>42,43</sup> To the best of our knowledge, this is the first attempt to derive an AA matrix solely from residue contact potentials. Because interresidue contacts are essential to maintain a protein fold, the contact propensity is expected to contain important information about similarity of amino acids related to their ability to conserve a fold. Calculated AA matrices were benchmarked on a set of structurally aligned protein sequences prepared for three levels of the sequence similarity, namely, the family, the superfamily, and the fold level. The performance of the newly derived AA matrices is compared with that of the existing matrices taken from AAIndex database.<sup>44</sup> Our AA matrices perform among the best in the benchmark set of fold level similarity. We observed that most of the tested AA matrices perform comparably on the alignment sets of family level, but the difference in performance becomes larger when the sequence similarity level becomes more distant. Interestingly, our AA matrices further outperform the existing matrices when suboptimal alignments are considered. It is notable that our matrices have little resemblance to the existing AA matrices, suggesting that there is another

local spot in the AA matrix space that is good for aligning sequences from the family level to the fold level.

## MATERIALS AND METHODS

This section is organized as follows. First, we describe methods to calculate the novel contact potential-based AA matrices. This first subsection is further divided into three paragraphs that describe the dataset used to derive knowledge-based amino acid pairwise contact potentials, the methods used to derive the contact potentials, and the methods employed to derive the AA matrices from the contact potentials. Then, we list the existing AA matrices used to compare with our newly derived contact potential-based AA matrices. Next, we describe the benchmark databases of sequence alignments used to evaluate the performance of the AA matrices. Finally, we explain the method used to calculate suboptimal alignments, the criteria to evaluate the performance of the AA matrices, and the method used to optimize gap penalties for each AA matrix.

### Derivation of Amino Acid Pairwise Contact Potentials

A series of AA matrices are derived from amino acid pairwise contact potentials. The pairwise potentials are derived from a representative set of 3156 protein structures. This representative set is selected by the PDB-REPRDB server with the default parameter set<sup>45</sup> that excludes proteins that have 30% or greater sequence identity with the other members in the set.

By the following methods described by Skolnick et al.,<sup>43</sup> we derived six different pairwise potentials. Six different potentials originate from the various threshold value combinations to define a residue pair contact (either 4.5 or 6.0 Å for any pair of heavy atoms from two amino acid residues), ways to calculate the reference state, and ways to calculate the average potentials from a library of structures. Two reference states used are the Gaussian random chain model, which considers the connectivity of a Gaussian polymer chain,<sup>46</sup> and the quasi-chemical approximation, where an expected number of contacts for a given amino acid pair is proportional to the fraction of both amino acids in a protein. Two ways used to average potentials include the partial composition correction method, which uses the average amino acid compositions of individual proteins among those in the library to derive a potential, and the composition correction method, where potentials calculated for individual proteins are averaged among those in the library to produce a final potential. The six derived amino acid contact potentials are summarized in Table I. In addition to the six newly derived potentials, we used a contact potential taken from Table II of the Skolnick's article,<sup>43</sup> termed here REFPG. This potential was derived in the same way as the PG, which uses 4.5 Å as the threshold distance of residue contacts, the partial composition averaging method and the Gaussian chain as the reference state, but with the much smaller structure library size (224 proteins) available at the time of the publication.

**TABLE I. Amino Acid Pairwise Contact Potentials Used to Derive AA Matrices**

Code	Threshold (Å)	Averaging Method	Reference State
CCPC	4.5	Composition Correction	Gaussian Chain
CC6PC	6.0		
CCPG	4.5	Partial Composition	Gaussian Chain
CC6PG	6.0	Correction	
CCPQ	4.5	Partial Composition	Quasi Chemical
CC6PQ	6.0	Correction	Approximation
REFPG <sup>a</sup>	4.5	Partial Composition	Gaussian Chain
		Correction	

<sup>a</sup>This potential is taken from Table 2 in the paper by Skolnick *et al.*<sup>43</sup>

The derived contact potentials are similar to each other; indeed, the correlation coefficient of the four potentials that use the contact threshold value of 4.5 Å ranges from 0.998 (PQ and PG) to 0.86 (PC and REFPG). We also tested the performance of the four potentials on a gapless threading test<sup>43</sup> using 100 template proteins. The four potentials again show similar performance; PC successfully recognized 97 native proteins, PQ and PG recognized 95 proteins, and REFPG recognized 94 proteins.

### Derivation of the AA Matrices from the Pairwise Potentials

The main idea presented in this article is to derive AA matrices from the pairwise contact preference of amino acids, which are described in the form of residue contact potentials. Corresponding positions in a pair of proteins of the same fold are considered to have a similar “environment,” which includes a similar contact pattern with residues surrounding the positions. Therefore, two residues sitting at equivalent positions in two proteins are expected to have a similar residue contact preference. For any two given amino acids, the similarity of the residue contact preference can be computed by comparing the residue contact potential values of the two to the 20 amino acids. One way to quantify the similarity is to simply compute the correlation coefficient of the contact potential values of the two columns of the two amino acids.

For each of the pairwise contact potentials listed in Table I, we derived an AA matrix, which contains the correlation coefficient (CC) of contact potential values. The matrices are named CC with the code of the potential, that is, CCPC is the AA matrix calculated from the potential PC, CC6PC from 6PC, CCPG from PG, CC6PG from 6PG, CCPQ from PQ, CC6PQ from 6PQ, and CCREFPG from REFPG. These seven AA matrices constitute the base series of our CC matrices.

### Optimizing CC Matrices

In addition to the base CC matrices described above, two different ways of optimization were applied to create even more matrices. The first optimized matrix series is calculated by the linear combination of the CCPC matrix with

the KOLA matrix,<sup>47</sup> which is one of the existing matrices used for comparison. We chose the KOLA matrix for the combination because it performs well in aligning sequences at fold level similarity (see Result section) and reflects the propensity of the main-chain torsion angles of amino acids, thus considered to contain complementary structural information to our contact potential-based CC matrices. We chose CCPC matrix among the base CC matrix series because it performs the best among the base CC series at the fold level (see Result section). First, values in the KOLA matrix are scaled down so that all diagonal entries have a value of 1, which is the same as the base CC matrices. Nine new matrices, CCX, where X ranges from 10 to 90, are derived with a different weight given to CCPC. The value of the position ( $i,j$ ) in CCX matrix,  $CCX_{i,j}$ , is defined as follows:

$$CCX_{i,j} = (X/100)*CCPC_{i,j} + ((100 - X)/100)*KOLA_{i,j} \quad (1)$$

For example, the CC10 matrix is computed as the linear combination of 10% of CCPC and 90% of KOLA.

The second form of optimization is to further numerically optimize the CCPC matrix so that the resulting matrix performs better on a benchmark alignment dataset. We employ the downhill simplex algorithm,<sup>40,48</sup> which finds at least a local optimal solution to a problem. In the optimization we fix all diagonal values to the value of 1, resulting in 192 adjustable parameters that correspond to 190 nondiagonal values and the opening/extension gap penalties. The target function to be optimized is the weighted sum of the percentage of correctly aligned residues in an alignment within a dataset, WSPC:

$$WSPC = \sum_{i=1}^n w_i PC_i$$

$$w_i = \begin{cases} 1.0 & \text{when } PC_i \geq 50 \\ 0.5 + 0.01 \times PC_i & \text{when } PC_i < 50 \end{cases} \quad (2)$$

where  $n$  is the number of alignments in the dataset,  $w_i$  is the weight given to a predicted alignment, and  $PC_i$  is the percentage of correctly aligned residues in the predicted alignment.



We optimize the CCPC matrix using three alignment datasets, namely, the family, superfamily, and fold level alignment benchmark sets of Lindahl and Elofsson's database (see below). This procedure produces three extra matrices, Fam01, Sfam01, and Fold01.

In what follows, we term the group of newly derived AA matrices the CC series, which includes the base CC matrices, CCX matrices and Fam01, Sfam01, and Fold01.

### Existing AA Matrices

Among 83 AA matrices stored in the AAIndex database<sup>44</sup> we choose the following 12 matrices for comparison. These 12 matrices are selected because they are either aimed to align distantly related sequences or expected to be able to align distantly related sequences because they reflect the structural information of proteins. The BLOSUM45 matrix is also included because the BLOSUM series is the most frequently used matrix group for sequence alignments and database searches.<sup>35,49,50,35</sup> Among the BLOSUM series, we choose BLOSUM45 because it is the matrix used for distantly related sequences in BLAST<sup>50</sup> and PSI-BLAST.<sup>35</sup> In addition to the 12 matrices, a random amino acid similarity matrix is generated. In the random matrix, numbers between -5 and 15 are randomly assigned. The 12 selected matrices are listed below. The ID we used in this study is followed by the accession code in the AAIndex database.

#### **BLAJ (BLAJ010101)**<sup>51</sup>

This matrix is constructed from the structural superimposition of 637 homologous domain pairs taken from the CATH database. The sequence identity of the domain pairs is less than 30%. This is the BC0030 matrix shown in Table II of the original article.

#### **BLOSUM45 (HENS920101)**<sup>52</sup>

This commonly used AA matrix is constructed from a set of ungapped block (alignment) sequences clustered with the sequence identity threshold value of 45%, that is, contribution to the amino acid frequency by homologous sequences sharing no less than 45% is averaged.

#### **JOHM (JOHM930101)**<sup>53</sup>

This matrix is derived based on structural alignments of 235 proteins. The majority of the sequence identity of the alignments falls between 15 and 40%.

#### **KOLA (KOLA920101)**<sup>47</sup>

Kolaskar and Kulkarni-Kale calculated an AA matrix from the similarity of the distribution of the ( $\phi$ ,  $\phi$ ) angles of the amino acids in 102 protein structures. This matrix is shown in Table II of the original article.

#### **KOSJ (KOSJ950115)**<sup>54</sup>

This AA matrix is numerically optimized so that it gives the highest probability of producing currently available descendant sequences given a phylogenetic tree.

#### **MIYS (MIYS930101)**<sup>55</sup>

The probability of codon substitutions is estimated considering energy increments (instabilities) upon amino acid exchange in protein structures.

#### **OVEJ (OVEJ920101)**<sup>39</sup>

A dataset of protein structure alignments of 34 homologous families is used to calculate the amino acid substitution pattern for this matrix.

#### **PRLA 1 (PRLA000101)**<sup>38</sup>

This matrix is derived from 122 protein structure alignments that do not have apparent sequence similarity. The formalism of Henikoff<sup>66</sup> is used to calculate the matrix.

#### **PRLA 2 (PRLA000102)**<sup>38</sup>

This matrix is derived by the authors of PRLA1 from 77 protein structure alignments that fall into the same homology level in the CATH database.

#### **QU\_C1 (QU\_C930101)**<sup>56</sup>

The correlation coefficients of so-called spatial preference factors of amino acid residues considering main-chain atoms are computed. The spatial preference factor shows how much atoms of two residues prefer to sit in one another's vicinity in a protein structure compared to the reference.

#### **QU\_C2 (QU\_C930102)**<sup>56</sup>

Created by the same procedure as QU\_C1, with the correlation coefficients of the spatial preference factors of side-chain atoms of amino acid residues also considered.

#### **QUIB (QUIB020101)**<sup>40</sup>

The GCB matrix,<sup>57</sup> which is derived from evolutionarily related protein sequence pairs, along with gap penalties, is optimized by the downhill simplex method to minimize the average root mean square deviation (RMSD) of aligned proteins in benchmark databases.

### Benchmark Databases of Pairwise Alignments

Primarily we use Lindahl and Elofsson's database<sup>58</sup> (L-E database) to benchmark AA matrices. In total it contains 1130 representative sequences taken from PDB,<sup>59</sup> each classified as a fold, superfamily, and family in hierarchical fashion according to the SCOP<sup>60</sup> database. We construct sets of pairwise sequence alignments in the above three levels of similarity. The pairs of sequences in the family level share the same family group in the SCOP database. Those in the superfamily level share the same superfamily group but belong to a different family, and those in the fold level share the same fold group but belong to a different superfamily. The proteins are aligned by CE, a protein three-dimensional structure alignment program.<sup>61</sup> This results in 2761 alignments at the fold level, 1395 alignments at the superfamily level, and 1076 alignments in the family level. The distribution of the sequence alignments in the three levels is shown in Figure 1(A). As can be seen in Figure 1, the sequence identity of the

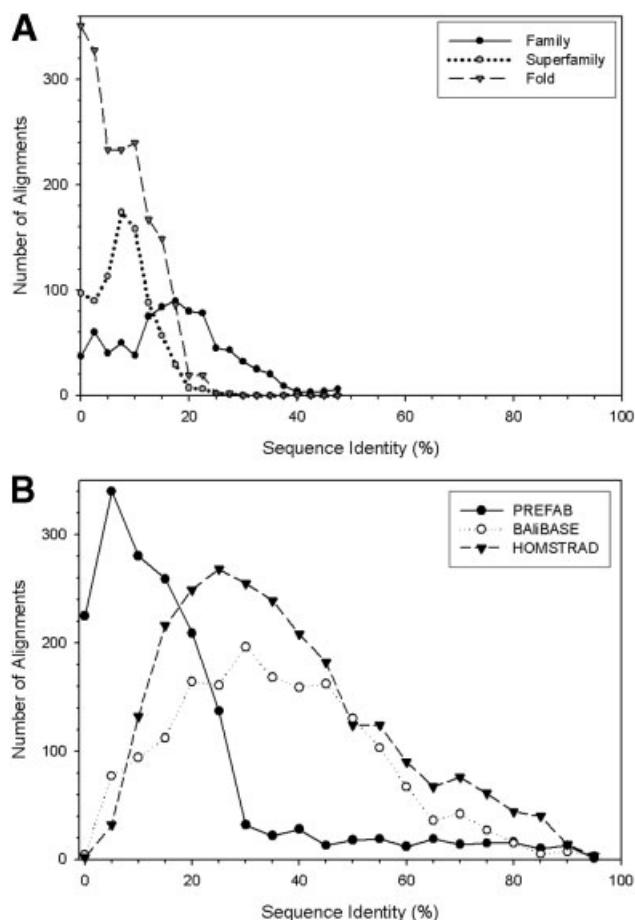


Fig. 1. The distribution of the sequence identity of the sequence alignment benchmark datasets. (A) The family, super family, and fold level alignment sets of the L-E database; (B) Prefab, BALiBASE, and HOMSTRAD database.

alignments is lower than that of the usual alignments benchmark sets. For example, the average sequence identity, even for the family level, is lower than 20%. Thus, these sets are the hardest benchmark sets in the study. The gap penalties for each AA matrix are independently optimized for the three levels of the L-E database.

In addition to the three sets of the L-E database, we use three other alignment benchmark databases, namely, BALiBASE,<sup>62</sup> PREFAB 4.0,<sup>63</sup> and HOMSTRAD.<sup>64</sup> BALiBASE consists of 165 reference alignments each containing at least two sequences. We extract the first two sequences from each of the multiple sequence alignments. PREFAB 4.0 consists of 1682 families each containing exactly two sequences supplemented with reference structural alignments. The HOMSTRAD database consists of curated multiple alignments of 1032 families originating from protein structure alignments, which contain at least two sequences. From each of the multiple sequence alignments in HOMSTRAD, the first sequence and one that has the least sequence identity to the first are selected to construct a pairwise alignment. The distribution of the sequence identity of these databases [Fig. 1(B)] shows that BALiBASE and HOMSTRAD have higher sequence iden-

tity than even the family level set in the L-E database. The peak of the distribution of the Prefab database is below 10%, but the long tail to the high sequence identity region pushes the average sequence identity to 19.7%, which is much higher than that of the superfamily level of the L-E database (7.7%).

The AA matrices including the CC series and those used for comparison, as well as the benchmark databases used, are available from our Web site, <http://dragon.bio.purdue.edu/aamatrices/>.

### Optimal and Suboptimal Alignment Algorithms

The global alignment algorithm<sup>65</sup> with free ending gaps and affine gap penalties is used to obtain the optimal alignment. As for suboptimal alignments, the method described by Saqi and Sternberg was implemented.<sup>66</sup> To obtain a next best alignment, this suboptimal alignment algorithm penalizes the current path in the dynamic programming matrix by subtracting a score of  $\Delta$  along the current path.  $\Delta$  equals 10% of the best score in the AA matrix used. This step is repeated  $(x - 1)$  times to obtain  $x$ -suboptimal alignment.

### Evaluating Alignments

A computed alignment is evaluated by the fraction (%) of correctly aligned residue pairs allowing a  $\pm 2$  residue shift. The reason of allowing the shift is because this small amount can be corrected by a comparative modeling method<sup>67</sup> when the alignment is used initially between the query sequence and the template.

### Optimizing Gap Penalties

For a given AA matrix, the optimal set of opening and extension gap penalties is searched in a hierarchical fashion. First, the opening and extension gaps are searched exhaustively with a coarse interval. Then, a finer interval size is used recursively around the optimal gap penalties identified by a previous coarse search. The gap penalty set is searched independently for the fold, superfamily, and family level set in the L-E database.

Concretely, the initial search round starts from an upper bound of the opening gap penalty ( $g_1^{up}$ ) of 100. The opening gap penalty ( $g$ ) is decreased from  $g_1^{up}$  by a step size ( $s_1$ ) of 10 until it reaches 0. For each  $g$ , different values of the extension gap penalty ( $e$ ) are tried in the descendant order from  $g$  to 0 with an interval of  $s_1$ . The new set of gap penalties is accepted (1) if the number of alignments that have 50% or more of the positions correctly aligned ( $N_{50}$ ) increases and the number of alignments with no correctly aligned positions ( $N_0$ ) decreases at the same time; (2) if  $N_{50}$  increases with a sacrifice of an increase of  $N_0$  up to 10% of  $N_{50}$ ; or (3) if  $N_0$  decreases with a sacrifice of a decrease of  $N_{50}$  up to 10% of  $N_0$ .

The resulting pair of the opening and extension gap penalties found in the initial search is termed ( $g_1, e_1$ ). Then, in the second round, the upper limit of the opening gap penalty ( $g_2^{up}$ ) is set to be  $g_1 + 25$ , and a finer step size ( $s_2$ ) of 5 is used in the same way to identify a good set of gap penalties, ( $g_2, e_2$ ). This step is further repeated three more

**TABLE II. Optimized Opening and Extension Gap Penalties for the Three Levels of the L-E Database**

Similarity Matrices	Family		Super Family		Fold	
	Opening	Extension	Opening	Extension	Opening	Extension
[BA] <sup>a</sup> BLAJ	24.5	3.9	18.1	2.5	8.9	4.2
[BO] BLOSUM45	6.5	2.9	6.0	1.5	2.2	2.1
[JO] JOHM	5.8	5.1	4.5	3.7	2.2	1.9
[KL] KOLA	15.1	1.1	10.7	1.0	7.8	2.9
[KS] KOSJ	86.1	29.6	98.9	17.7	60.6	0.1
[MY] MIYS	0.4	0.4	0.3	0.3	0.3	0.1
[OV] OVEJ	3.8	3.6	3.5	1.8	1.6	1.4
[P1] PRLA 1	3.3	1.2	3.3	0.9	1.9	0.8
[P2] PRLA 2	11.5	2.6	11.0	2.1	4.0	3.3
[Q1] QU_C 1	1.0	0.1	1.4	0.1	0.9	0.1
[Q2] QU_C 2	1.8	0.1	1.1	0.1	1.2	0.1
[QB] QUIB	9.6	0.7	8.9	0.5	2.7	2.5
[CC] CCPC	0.8	0.1	0.5	0.1	0.2	0.1
[CG] CCPG	0.7	0.1	0.7	0.1	0.2	0.1
[CQ] CCPQ	0.8	0.1	0.8	0.1	0.3	0.1
[CR] CCREFPG	0.8	0.1	0.6	0.1	0.5	0.1
[6C] CC6PC	0.6	0.1	0.4	0.1	0.1	0.1
[6G] CC6PG	0.9	0.1	0.6	0.1	0.4	0.1
[6Q] CC6PQ	0.6	0.1	0.5	0.1	0.3	0.1
[C1] CC10	1.4	0.1	0.7	0.1	0.4	0.4
[C2] CC20	1.2	0.1	0.9	0.1	0.5	0.3
[C3] CC30	1.3	0.1	0.7	0.1	0.3	0.2
[C4] CC40	0.9	0.1	0.4	0.4	0.4	0.1
[C5] CC50	0.5	0.1	0.8	0.1	0.5	0.1
[C6] CC60	0.4	0.1	0.8	0.1	0.3	0.2
[C7] CC70	0.5	0.1	0.8	0.1	0.2	0.1
[C8] CC80	0.8	0.1	0.5	0.1	0.4	0.1
[C9] CC90	0.5	0.1	0.6	0.1	0.2	0.2
[FA] Fam01	0.8	0.1	- <sup>b</sup>	-	-	-
[SF] Sfam01	-	-	0.8	0.1	-	-
[FO] Fold01	-	-	-	-	0.8	0.1
[RD] RANDOM	5.0	4.2	23.8	0.1	21.8	0.3

<sup>a</sup>The code in the parenthesis corresponds to those used in Fig. 4–8.

<sup>b</sup>Gap penalties for super family and fold level are not shown here because Fam01 matrix and its associated gap penalties are optimized for family level alignments of the L-E database. The identified gap penalties for the family level are used also for the superfamily and the fold level alignments. For the same reason, the family level and the fold level gap penalties for Sfam01, the family level and the super family level gap penalties for Fold01 are not shown.

times with the following conditions:  $(g_3, e_3)$  determined by  $g_3^{up} = g_2 + 10$  and  $s_3 = 1$ ;  $(g_4, e_4)$  is determined by  $g_4^{up} = g_3 + 5$  and  $s_4 = 0.5$ ; and finally  $(g_5, e_5)$  is determined by  $g_5^{up} = g_4 + 1$  and  $s_5 = 0.1$ . The resulting gap penalties,  $(g_5, e_5)$ , are summarized in Table II for each sequence similarity level. For the majority of the AA matrices, the opening gap penalty decreases as the average sequence identity of the alignments decreases (from family level to fold level). In Figure 2, the CCPC matrix (Table III) is compared with BLOSUM45.

## RESULTS

### Comparison of the CC Matrices to the Other Matrices

To begin with, we have compared the CC matrix series with the other AA matrices. In Figure 2, the CCPC matrix is compared with BLOSUM45. Interestingly, these two matrices are quite different, the correlation coefficient of these two matrices is only 0.54, indeed. Figures 2(A) and

(B) show the clustering results of amino acids based on the BLOSUM45 and CCPC matrices, respectively. Basically, hydrophobic amino acids form a separate branch from the others in both trees, although there are some differences. In the tree for the CCPC [Fig. 2(B)], the branch for hydrophobic amino acids (the upper branch with A to C) is clearly distinctive from the rest. Alanine and Cysteine are included in the branch of hydrophobic residues in CCPC, but not in BLOSUM45 (the branch of I to W). Interestingly, Glutamine (Q) and Glutamic Acid (E), Aspartic Acid (D), and Asparagine (N) are not paired in the CCPC, as they are in the BLOSUM45. In Figure 2(C), values in the CCPC matrix are directly compared with those of BLOSUM45. A moderate correlation between the two matrices is illustrated in this plot: all the amino acid pairs with a positive BLOSUM45 score have a CCPC score of 0.5 or higher, and when the BLOSUM45 score is higher than 4, all the amino acid pairs have a CCPC value of 1.0. These amino acid pairs with a CCPC value of 1.0 are pairs of

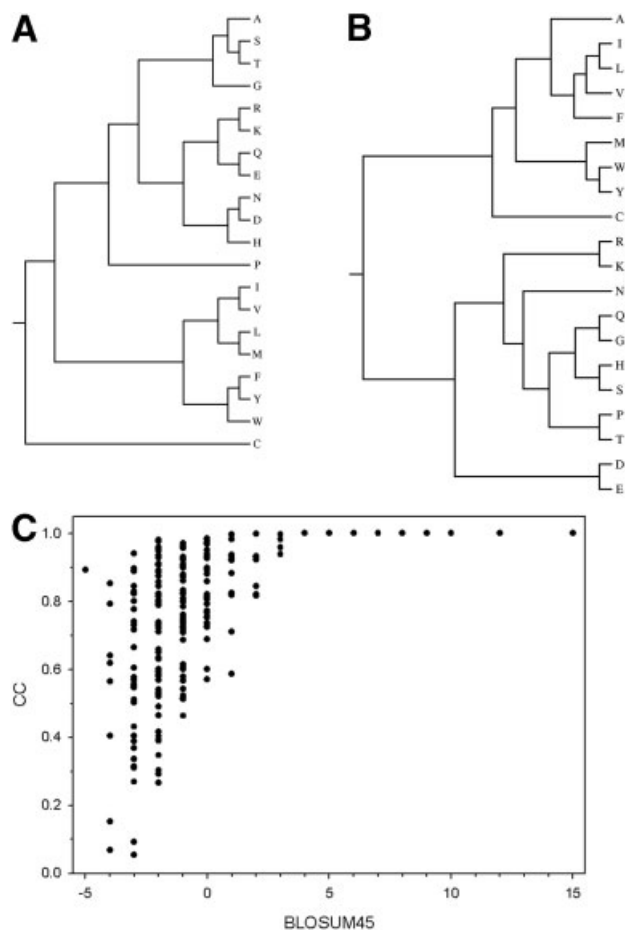


Fig. 2. Comparison of the CCPC matrix with BLOSUM45. (A, B) Clustering of amino acids using values in AA matrices as the pairwise distance of the BLOSUM45 and CCPC matrix, respectively. The distance of a pair of amino acid is defined as the opposite sign of the corresponding value of the amino acid in the AA matrix (this is to have a closer distance for a pair with a high positive value in the AA matrix). The Neighbor program in the PHYLIP package<sup>71</sup> with the UGPMA method was used to draw the trees. (C) Values in the CCPC matrix relative to those of BLOSUM45. Each point represents an individual amino acid pair, with the x coordinate being the value in BLOSUM45 of that amino acid pair and the y coordinate being that of the CCPC matrix.

identical amino acids (note that the diagonal values of BLOSUM45 differ from one amino acid to another). On the other hand, there are many pairs of amino acids that have a negative value in BLOSUM45 but have a high value in the CCPC matrix. Those examples include pairs of Tryptophan with Cysteine (−5, 0.982), Proline with Cysteine (−4, 0.852), and Tryptophan with Serine (−4, 0.792). The numbers in parentheses show the values for the pairs in the BLOSUM45 and CCPC matrices, respectively. The weak correlation between the two matrices illustrates that the pairwise amino acid residue contact propensity is one factor that governs the amino acid mutation ratio in proteins, but there are certainly different factors involved in mutation events.

We further compare all the AA matrices by computing their correlation coefficients between one another (Fig. 3). KOSJ (column 5) and the random matrix (clm. 32) are very

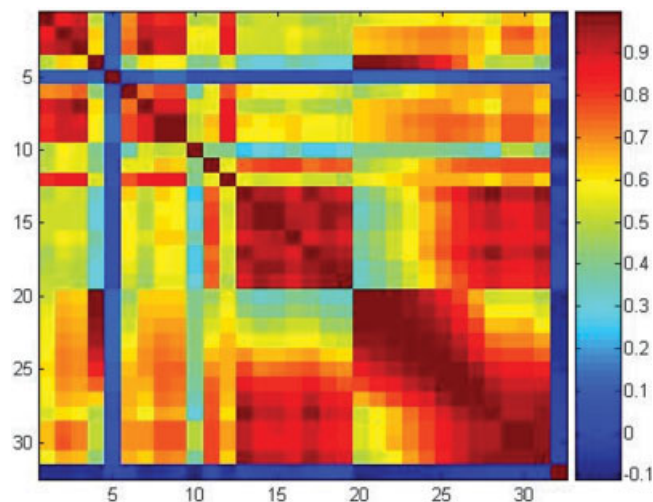


Fig. 3. The comparison of the AA matrices used in this study. The correlation coefficient of each pair of the AA matrices is shown in a color code. The AA matrices (1–32) are ordered in the same way as Table II. The color code ranges from dark blue for −0.1 to dark red for 1.0.

different from the others. Roughly speaking, we observe three clusters in this diagram: the first cluster, which is on the upper left corner, is comprised of BLAJ, BLOSUM45, JOHM, MIYS, OVEJ, PRLA1, PRLA2, and QUIB. The majority of these matrices are created based on a set of multiple sequence alignments. QUIB is a numerically optimized matrix, but it has a high correlation to the matrices in the first cluster. This is because the initial values of the matrix are the GCB matrix,<sup>57</sup> which was derived from alignments of homologous protein sequences. The fact that matrices in this cluster have a high correlation to BLOSUM45 indicates that these heavily reflect the mutation rate of amino acids in evolutionarily related protein sequences. Below we call these matrices in the first cluster homology based. The second cluster, which is in the middle of the diagram, is comprised of a series of matrices calculated from a knowledge-based pairwise potential (from CCPC to CC6PQ, clm. 13 to 19). It is noteworthy that this cluster has little correlation with the first cluster. Interestingly, QU\_C2 (clm. 11) has a good correlation with this second cluster. Indeed, QU\_C2 is the correlation coefficient of the “spatial preference factor” of amino acids considering side chain atoms, which essentially considers the contact propensity of amino acid pairs. Interestingly, QU\_C1, which is computed in the same way as QU\_C2 but considers main-chain atoms, has a good correlation with the first cluster but not with the CC series. The third cluster on the right bottom corner (clm. 20 to 31) contains the modified matrices from the original CC matrix, the series of the mixture of CC with KOLA (clm. 4), with a different ratio (CC10 to CC90) and further optimized matrices (Fam01, Sfam01, and Fold01). Therefore, it is natural that these matrices have a high correlation with the second cluster and also with KOLA. To conclude, the CC matrices are not slight modifications of conventional matrices such as BLOSUM45, but are instead located in a completely different spot of the matrix space.



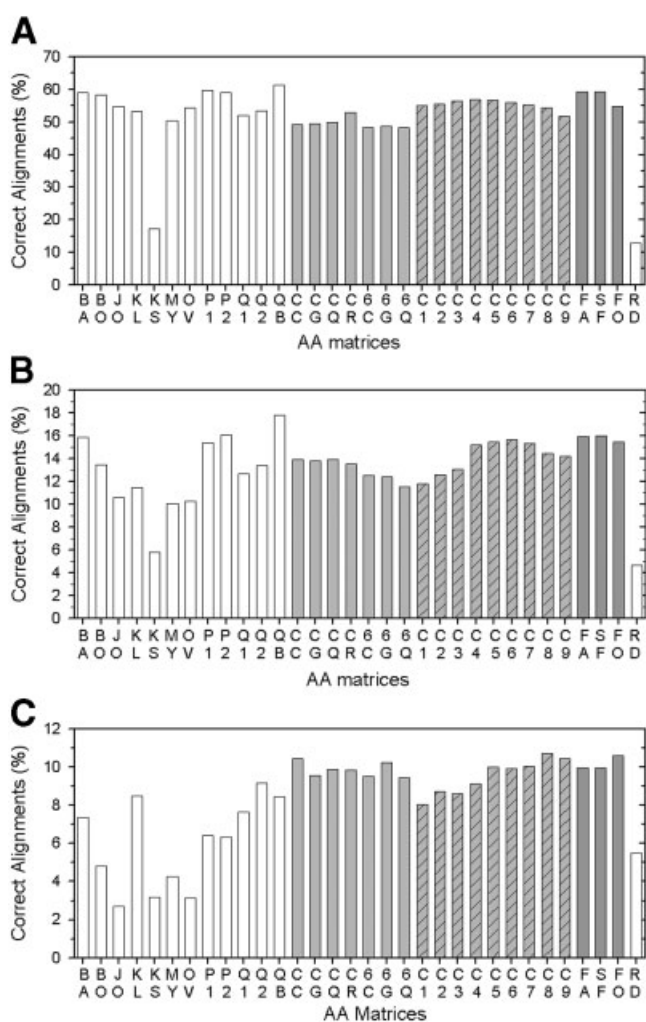


Fig. 4. The fraction of correct alignments computed with different AA matrices. An alignment is counted as correct here when more than 50% of the positions in the alignment match within  $\pm 2$  residue shifts, with the reference alignment prepared by by CE program. The three benchmark sets of a different similarity level in the L-E database are used. (A) Family level set; (B) superfamily level; and (C) fold level. The abbreviation code for matrices is shown in Table II. White bars on the left, existing matrices taken from AAIndex database; light gray bars, the base CC matrices; hashed bars, the CCX series; dark gray bars, optimized CC series; the white bar on the most right, the random matrix.

### Performance on the L-E Database

Figure 4 shows the performance of the matrices on the three levels of benchmark alignments in the L-E database. For each AA matrix, the gap penalties optimized for each level of the alignment set are used. The y-axis shows the fractions of alignments in the data set that match more than 50% to the correct alignment, allowing  $\pm 2$  residue shifts.

In the family level benchmark set [Fig. 4(A)], homology-based matrices perform relatively well compared to their performance on the superfamily and the fold level sets. The best-performing matrix is QUIB (61.3%). PRLA1 (59.8%), Fam01 (59.3%), Sfam01 (59.3%, same as Fam01), and PRLA2 (59.1%) follow in this order. BLOSUM45's

performance is also not bad (the seventh rank; 58.4%). Among the top 10 performing matrices, five are homology-based and another five come from the CC series. It is natural that the homology-based matrices perform well at the family level. What is more interesting is that the CC matrices also perform comparably, considering their very different nature. At the superfamily level [Fig. 4(B)], QUIB still perform the best, but the performance of the homology-based matrices is lower relative to the CC series. Six out of the top 10 performing matrices are now from the CC series, and BLOSUM45 drops in rank to 18th place.

It is remarkable that at the fold level [Fig. 4(C)] the majority of the matrices in the CC series clearly outperform existing matrices. Indeed, the top 10 matrices all come from the CC series. The top five matrices are CC80 (C8), Fold01 (FO), CC90 (C9), CC (CC), and CC6PG (6G). Note that at this level of sequence identity, BLOSUM45 performs no better than the random matrix (RD). Among the existing matrices tested (BA to QB), QU\_C2 (Q2), KOLA (KL), and QUIB (QB) show the top three best performances. As mentioned above, QU\_C2 essentially captures the pairwise contact propensity of amino acid side chains, as do the CC matrices. Thus, the good performance of QU\_C2 is consistent with the performance of the CC series. On the other hand, KOLA captures the propensity of dihedral angles of amino acids, and is thus based on structural information very different from the CC series (see its weak correlation to the CC series in Fig. 3). KOLA is unique because it also does not resemble the homology-based matrices.

Among the CCX series, which is the mixture of CC and KOLA, several matrices performed better than the original KOLA and CC, with CC80 performing the best of all [Fig. 4(C)]. These mixed matrices also performed well at the family and superfamily levels. Interestingly, the peak performance among the CCX series shifts from matrices with less CC content to ones with more as the sequence identity level of the benchmark set drops. The peak at the family level is CC40 [Fig. 4(A)], goes to CC60 at the superfamily level [Fig. 4(B)], and finally to CC80 at the fold level [Fig. 4(C)]. This observation may imply that residue contact propensity at each position of a structure is one of the ultimately conserved properties of protein fold.

Figure 5 shows more detailed distribution of the correctly aligned residues in the datasets for eight matrices. BLAJ and BLOSUM45 are chosen as the representative of the homology-based matrices. KOLA and QU\_C2 are chosen from the existing matrices and CC, CC08, Fold01 are chosen from the CC series because of the good performance at the fold level [Fig. 4(C)]. In addition, the random matrix is chosen to draw the baseline. BLAJ and BLOSUM45 show superior performance at the family level [Fig. 5(A)]. At the superfamily level, all the matrices show similar performance except for the random, BLOSUM45 and KOLA, although the three CC series matrices have slightly better performance in producing good alignments with 70% or higher correctly aligned positions. Finally in the fold level, the three CC series matrices outperform the other matrices.



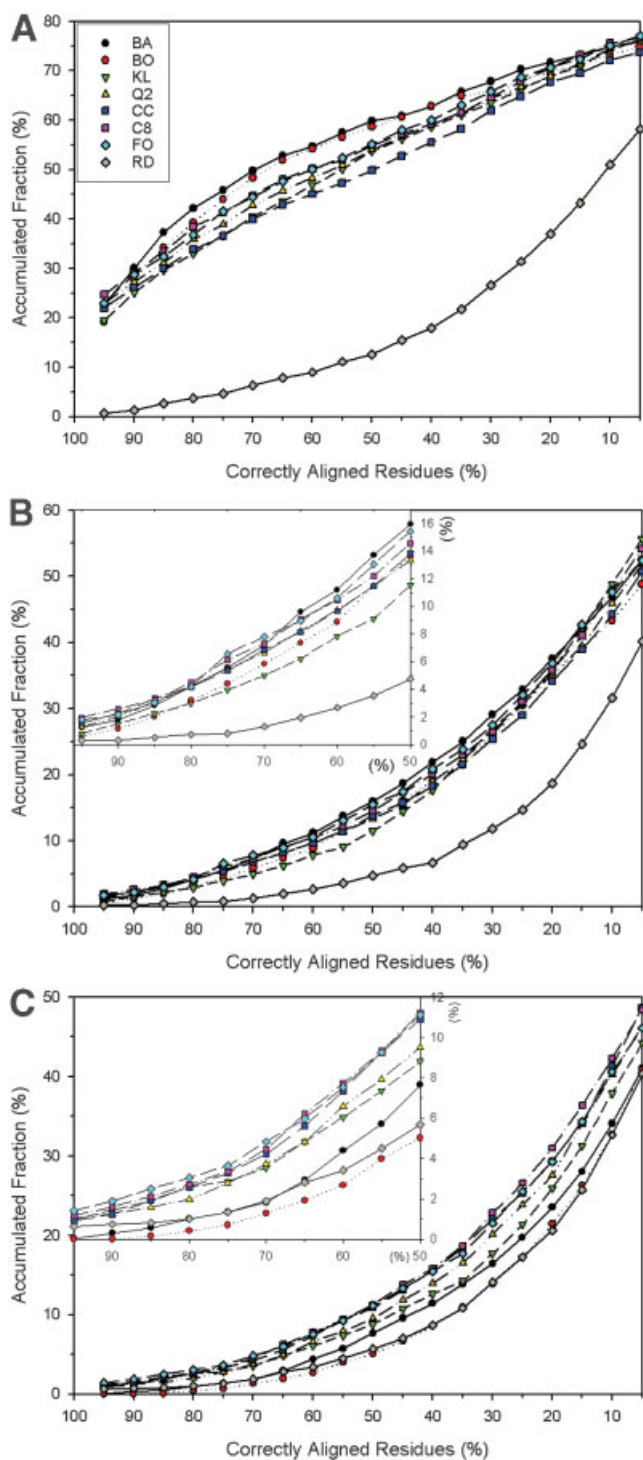


Fig. 5. The distribution of fraction of correctly aligned residues within  $\pm 2$  residues. Black, BLAJ; red, BLOSUM45; green, KOLA; yellow, QU\_C2; blue, CC; pink, CC80; light blue, Fold01; gray, random matrix. (A) Family level; (B) superfamily level; (C) fold level. For the superfamily and fold level, a zoomed plot for the x-axis of 100 to 50 is inserted. The y value of a particular x shows the accumulated fraction of alignments whose x to  $x + 5$  (%) of positions are aligned correctly.

We have also used a different measure of the alignment accuracy, that is, the scaled normalized distance ratio (SNDR)<sup>68</sup> (Fig. 6, Table IV). SNDR measures the average

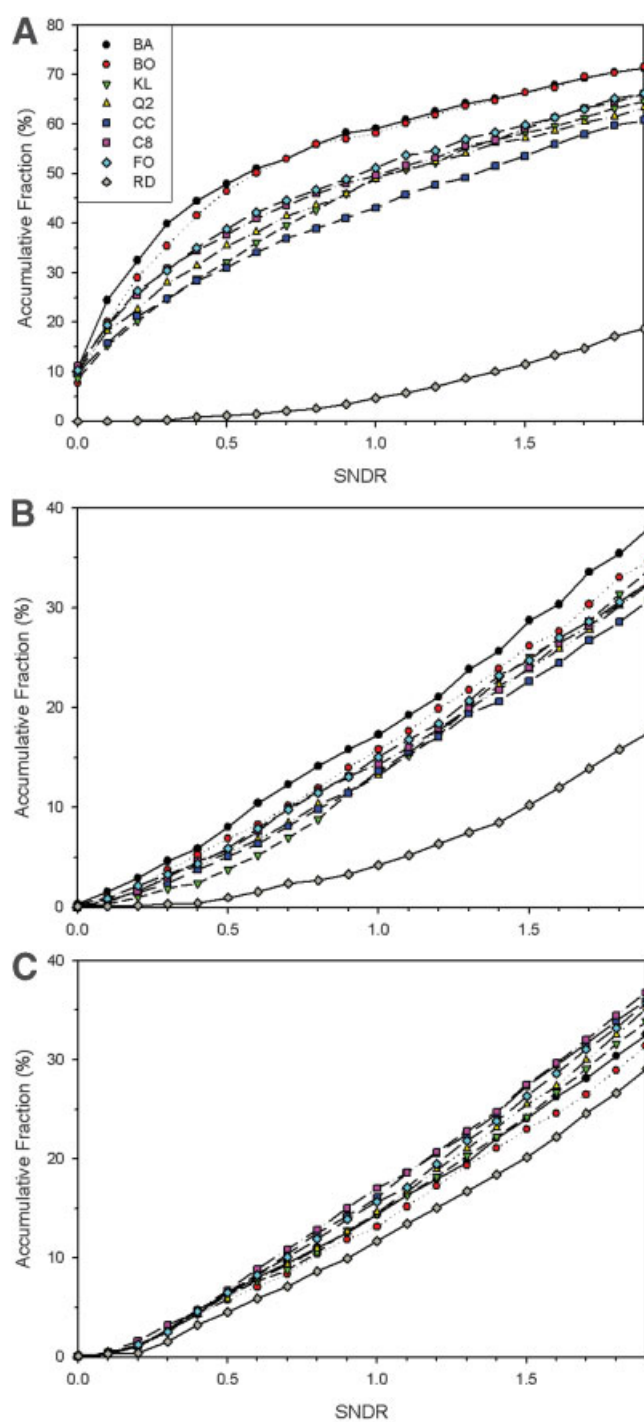


Fig. 6. The distribution of the scaled normalized distance ratio (SNDR) proposed by Blake & Cohen (2001). The same color code is used as Figure 5: (A) family level; (B) superfamily level; (C) fold level.

of the ratio of corresponding inter residue distances ( $\text{\AA}$ ) of the correct alignment and a predicted alignment:

$$\text{SNDR} = \frac{1}{N} \sum_{i=1}^N \left| 1 - \frac{d_i^{\text{predicted}} + 1}{d_i^{\text{structure}} + 1} \right|, \quad (3)$$

TABLE III. CCPC matrix (upper half) and CC80 matrix (lower half)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	1.000	0.580	0.514	0.266	0.822	0.709	0.463	0.736	0.723	0.962	0.956	0.513	0.971	0.976	0.861	0.710	0.880	0.949	0.959	0.968
A	1.000	1.000	0.814	0.566	0.605	0.825	0.601	0.820	0.859	0.405	0.389	0.959	0.579	0.465	0.798	0.879	0.809	0.658	0.687	0.417
R	0.596	1.000	1.000	0.844	0.650	0.930	0.808	0.926	0.883	0.292	0.269	0.821	0.568	0.393	0.822	0.920	0.806	0.641	0.630	0.310
N	0.411	0.651	1.000	1.000	0.431	0.766	0.932	0.724	0.689	0.068	0.053	0.571	0.311	0.152	0.604	0.754	0.600	0.404	0.403	0.092
D	0.213	0.453	0.855	1.000	1.000	0.821	0.954	0.922	0.540	0.520	0.819	0.752	0.619	0.930	0.951	0.923	0.814	0.809	0.556	
C	0.658	0.484	0.520	0.345	1.000	1.000	0.739	0.724	0.314	0.303	0.586	0.490	0.369	0.727	0.792	0.719	0.576	0.594	0.335	
Q	0.699	0.840	0.744	0.613	0.622	1.000	1.000	0.935	0.565	0.546	0.789	0.793	0.665	0.929	0.939	0.910	0.845	0.825	0.576	
E	0.550	0.613	0.646	0.746	0.409	0.789	1.000	1.000	0.552	0.530	0.832	0.773	0.634	0.909	0.960	0.907	0.823	0.817	0.569	
G	0.589	0.656	0.741	0.579	0.663	0.763	0.591	1.000	1.000	0.998	0.336	0.930	0.983	0.728	0.526	0.760	0.886	0.899	0.997	
H	0.578	0.787	0.806	0.683	0.838	0.838	0.579	0.748	1.000	1.000	0.313	0.922	0.982	0.716	0.503	0.742	0.875	0.893	0.997	
I	0.770	0.324	0.234	0.054	0.584	0.432	0.251	0.452	0.442	1.000	1.000	0.523	0.388	0.727	0.873	0.785	0.584	0.614	0.347	
L	0.865	0.443	0.215	0.042	0.569	0.596	0.342	0.437	0.424	0.798	1.000	1.000	0.970	0.891	0.733	0.877	0.981	0.974	0.932	
K	0.542	0.947	0.657	0.557	0.439	0.835	0.649	0.631	0.666	0.269	0.430	1.000	1.000	0.800	0.600	0.799	0.937	0.938	0.984	
M	0.877	0.643	0.454	0.249	0.685	0.702	0.392	0.634	0.618	0.744	0.870	0.518	1.000	1.000	0.898	0.927	0.941	0.940	0.740	
F	0.781	0.552	0.314	0.122	0.742	0.627	0.295	0.532	0.607	0.786	0.886	0.410	0.908	1.000	1.000	0.923	0.792	0.789	0.542	
P	0.689	0.638	0.658	0.483	0.682	0.744	0.582	0.743	0.727	0.582	0.573	0.582	0.713	0.640	1.000	1.000	0.897	0.908	0.773	
S	0.568	0.803	0.868	0.735	0.780	0.761	0.634	0.751	0.948	0.421	0.402	0.798	0.586	0.580	0.718	1.000	1.000	0.983	0.888	
T	0.704	0.747	0.645	0.480	0.792	0.838	0.575	0.728	0.826	0.608	0.594	0.728	0.702	0.739	0.742	0.870	1.000	1.000	0.904	
W	0.759	0.658	0.513	0.323	0.814	0.751	0.461	0.676	0.758	0.709	0.800	0.467	0.917	0.930	0.753	0.734	0.818	1.000	1.000	
Y	0.767	0.650	0.504	0.322	0.776	0.747	0.475	0.660	0.786	0.719	0.714	0.491	0.779	0.930	0.752	0.763	0.906	0.886	1.000	
V	0.774	0.334	0.248	0.074	0.590	0.445	0.268	0.461	0.455	0.978	0.798	0.278	0.746	0.787	0.592	0.434	0.618	0.710	0.723	1.000

TABLE IV. SNDR. A. The Distribution of SNDR of the Three Levels of the L-E Database

	Family		Superfamily		Fold	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
BLAJ	1.449	1.925	2.533	1.526	2.663	1.580
BLOSUM45	1.467	1.859	2.668	1.554	2.671	1.612
KOLA	1.665	1.718	2.600	1.410	2.598	1.562
QU_C2	1.698	1.824	2.686	1.470	2.557	1.569
CC	1.825	1.853	2.739	1.480	2.489	1.503
CC80	1.614	1.793	2.623	1.447	2.474	1.511
FOLD01	1.583	1.762	2.640	1.461	2.528	1.571
RANDOM	3.556	1.903	3.158	1.371	2.710	1.668

The three best average SNDR values for each level are shown in bold.

TABLE IV B. The t-test Results of the Difference of the Distribution of the SNDR

	Family			Superfamily			Fold		
	CC <sup>a</sup>	C8	FO	CC	C8	FO	CC	C8	FO
BA	4.480	1.996	1.632	3.511	1.549	1.839	4.049	4.396	3.083
BO	4.332	1.805	1.433	1.194	0.772	0.473	4.201	4.544	3.242
KL	2.027	0.645	1.056	2.476	0.414	0.724	2.563	2.916	1.623
Q2	1.565	1.039	1.441	0.932	1.105	0.793	1.582	1.937	0.663
RD	20.792	23.66	24.248	7.553	9.769	9.392	4.995	5.33	4.037

SNDR distribution of CC, CC80, Fold01 are compared with BLAJ, BLOSUM45, KOLA, QU\_C2, and the random matrix by the t-test. The t-value of two distributions is shown. The color codes indicate that the two distributions are significantly different: red and yellow, the CC series matrix (CC, C8 or FO) is significantly worse than the compared one with the risk level of 0.05 and 0.1, respectively; blue and green, our matrix is significantly better than the compared one with the risk level of 0.05 and 0.1, respectively.

<sup>a</sup>The abbreviation code of the matrices is shown in Table 2.

where  $N$  is the number of aligned residue pairs which exist both in the reference structure-based alignment and in the predicted alignment,  $d_i^{\text{structure}}$  is the distance (Å) of the residue pair  $i$  in the structure alignment, and  $d_i^{\text{predicted}}$  is the distance of the residue pair  $i$  indicated by the predicted

alignment. At the family level, the performance of the CC and CC80 is significantly worse than most of the other matrices judging from the  $t$ -test results [Table IV(B)]. However, the superior performance of the three CC series matrices is shown to be statistically significant at the fold level.

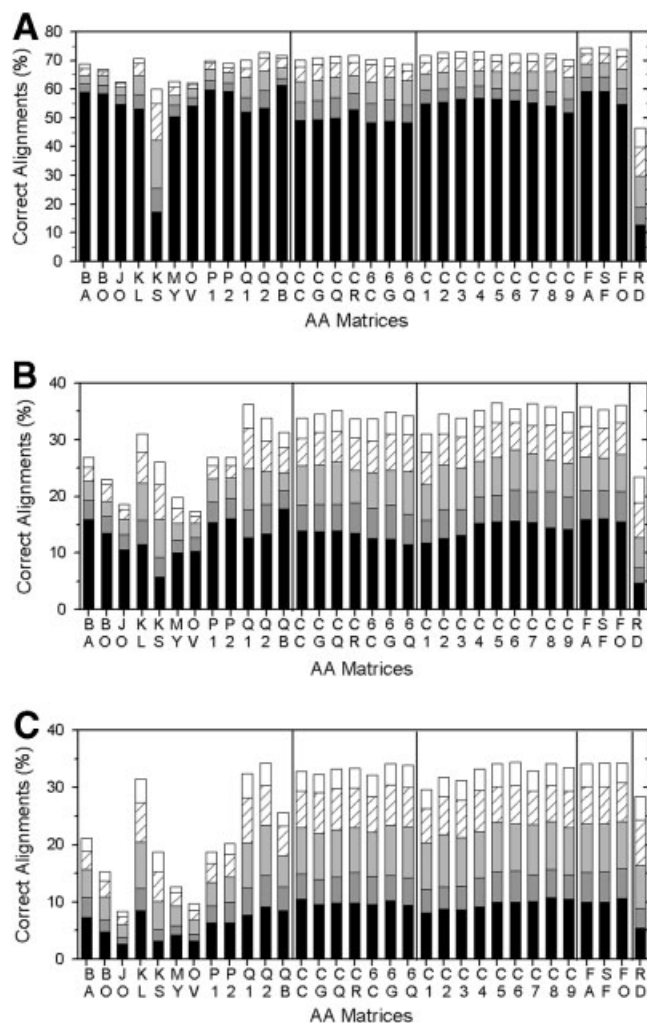


Fig. 7. Fraction of correct top scoring suboptimal alignments computed with different AA matrices. (A) Family level; (B) superfamily level; (C) fold level. Black, only the optimal alignment is considered; dark gray, a correct one is included among the top two scoring alignments; gray, up to the fifth suboptimal alignments are considered; hashed bars, up to 10 suboptimal alignments are considered; white, up to 15 suboptimal alignments are considered.

Note here that the performance of the matrices in terms of SNDR could be better if SNDR is integrated in the target function for optimizing the gap penalty in a certain way.

### Suboptimal Alignments

From a practical point of view, an algorithm is useful not only when it gives a good result with a top score, but also if a good result is included in the top  $n$ -th scoring results (if  $n$  is reasonably small). In the same spirit, the world-wide protein structure prediction competition, CASP (Critical Assessment of Techniques for Protein Structure Prediction), allows a prediction group to submit up to five prediction models.<sup>22</sup> In Figure 7, the performance of the matrices on the L-E database considering the top 1, 2, 5, 10, and 15 suboptimal alignments is examined. Interestingly, in all three levels of sequence identity, the performance of the CC matrices grows more than existing

matrices when suboptimal alignments are considered. At the family level [Fig. 7(A)], Fam01 and Sfam01 are ranked the top one and two matrices when two or more suboptimal alignments are counted. When the top 15 suboptimal alignments are considered, CC matrices occupy all five top ranks. It is even more striking that the CC matrices are dominant in the superfamily and the fold levels. This interesting feature of the CC matrices would be very valuable for template-based protein structure prediction. In the first step of a prediction process, 10 to 15 suboptimal alignments between a query sequence and a template structure can be pooled. Then, the most probable alignments can be further screened from the pool in the subsequent step by considering detailed structure matches between the query and the template.

### Performance on the Other Existing Benchmark Datasets

To confirm that the CC matrices also perform well in the other available benchmark databases of alignments, we have carried out the comparison of the AA matrices using the BALiBASE, PreFab, and HOMSTRAD databases. In this benchmark test, we used gap penalties optimized for the three levels of the L-E database (Table II), which are not specifically trained for this database.

The results in Figure 8 show that CC matrices have a comparable performance to the homology-based matrices. These graphs [Fig. 8(A–D)] show a profile similar to the one for the family level set of the L-E database [Fig. 4(A)]. This is reasonable because the distribution of the sequence identity of these databases is similar to that of the family level and superfamily level data set in the L-E database. The different gap penalties used do not affect much of the results in the majority of the cases.

To verify the validity of the performance measure used throughout this research, an alternative measure, the average sequence identity of all the alignments calculated by each AA matrix, is also used for these datasets. As an example, the results for BALiBASE are shown in Figure 8(B). The overall profile of Figure 8(B) is almost identical to Figure 8(A). The results for Prefab and HOMSTRAD using the two different performance measures are also closely similar to each other (data not shown). The close agreement of the results by the two different measures confirms that the conclusion drawn on the CC matrices is not just the artifact caused by the specific performance measure used, but capturing the positive characteristics of them.

### DISCUSSION

In this article we have reported our findings that the AA matrices simply derived from the correlation coefficients of amino acid pairwise contact potentials, the CC series, work well in protein pairwise sequence alignments. The CC matrices clearly outperform existing AA matrices when sequences to be aligned diverge, especially at the fold level similarity in the L-E database. Interestingly, the advantage of the CC series over the AA matrices increases when suboptimal alignments are considered. These charac-



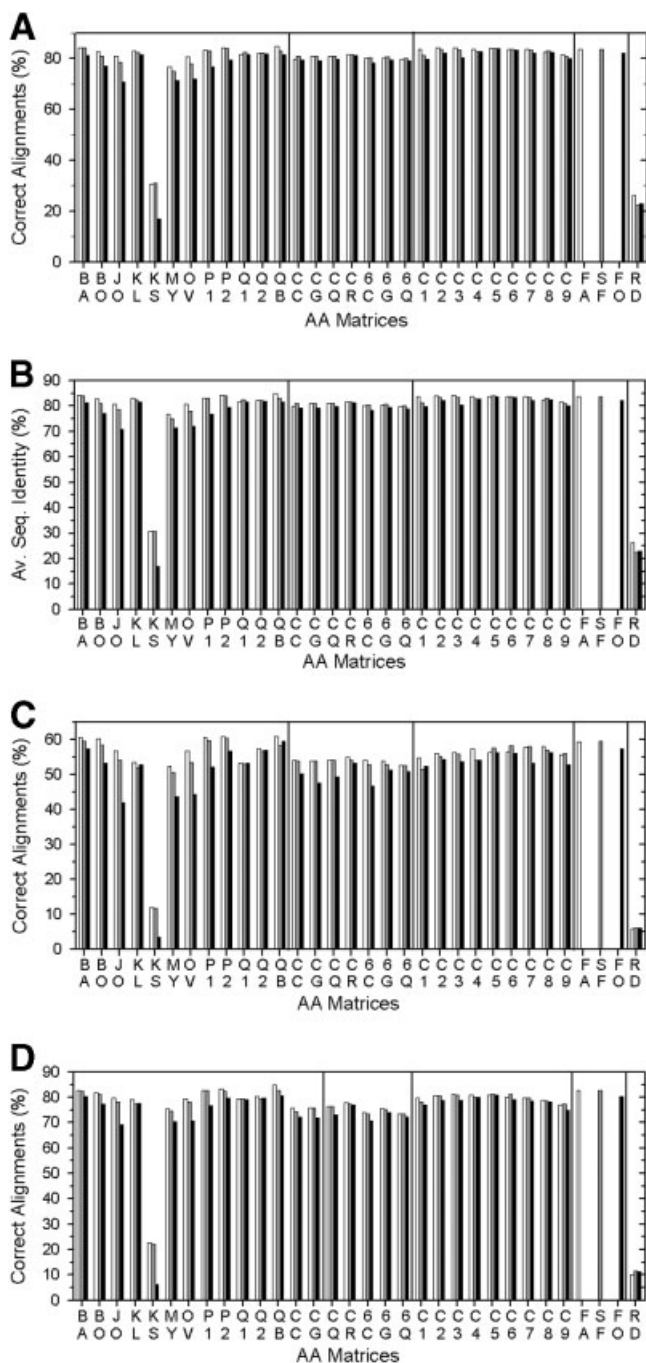


Fig. 8. The AA matrices are benchmarked on BALiBASE, Prefab, HOMSTRAD database. For each AA matrices, three different gap penalty sets are used. White, the gap penalties optimized for family level set in the L-E database; gray, gap penalties for the superfamily level; black, those for the fold level. Only one bar is shown in Fam01, Sfam01, and Fold01 because these three matrices are optimized for a specific level (see Table II). (A) BALiBASE with the y-axis being the fraction of the correct alignments; (B) BALiBASE with the y-axis being the average sequence identity of alignments in the dataset; (C) Prefab; (D) HOMSTRAD.

teristics of the CC matrices are very useful in protein structure modeling, because suboptimal alignments between the query sequence and template protein structures can be used as multiple starting points of modeling. In

template-based protein structure prediction/modeling, it is crucial that the correct alignment exist among the several top suboptimal alignments but not necessarily as the top-scoring alignment because spurious alignments could be eliminated during subsequent steps of structure optimization.

In Figure 4(A–C), it is interesting to see that the best performing matrix in the CCX series (gray hashed bars,  $X = 10–90$ ) differs as the dataset changes from the family to the fold levels. The more the sequence identity of the dataset drops, the more the fraction of the CCPC matrix is needed for the best performing CCX matrix. This fact implies that the amino acid dihedral angle propensity and the contact propensity are the two major factors that can be utilized for aligning distantly related protein sequences, but the contact propensity of positions in a protein structure is a property which is ultimately conserved in a protein fold.

Development of AA matrices for aligning distantly related protein sequences has been investigated over the years, but surprisingly, our CC matrix series is so far the only one that is derived solely from amino acid pairwise potentials. It is intriguing that the CC matrices are very different from existing homology-based matrices (Fig. 3) but still perform well when compared to closely related protein sequences and outperform them in fold level sequence similarity. This fact implies that the CC matrices locate to a different spot in the AA matrix space from the conventional homology-based AA matrices. It would be worthwhile to search the optimal AA matrices for distantly related protein sequences in the basin found by the CC series or to design novel AA matrices by considering amino acid residue contact propensity.

The importance of sequence alignment will continue to increase in this structural genomics era,<sup>69</sup> because more and more experimentally solved protein structures become available as templates for modeling. In a recent paper on a de novo (ab initio) high-resolution structure prediction, it is pointed out that one of the keys to a successful prediction is to bring a low resolution model to the vicinity of the optimal conformation,<sup>70</sup> because then a structure optimization using atomic-detailed potential could bring it to the correct model with an RMSD of 1–2 Å. The same strategy would be naturally applied for template-based structure prediction. Practically, the CC matrices could be used to make a better alignment between the query sequence and a template structure, once the template structure is identified by a certain threading method and if the template structure does not share a high-sequence similarity to the query. It would be also interesting to integrate the CC matrices into a threading or a database search algorithm, which is left for future work. To conclude, the contact potential-based AA matrices will open a new possibility in alignments of distantly related proteins that barely share sequence identity each other.

## ACKNOWLEDGMENTS

The authors thank Stan Luban for proofreading this manuscript and Yifeng Yang for making Figure 3. D.K.

acknowledges the support from the National Institute of General Medical Sciences of the National Institutes of Health (R01 GM-075004) and the Purdue Research Foundation. Y.H.T. was partially supported by Howard Hughes Summer Internship in 2005.

## REFERENCES

- Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 1998;282:740–744.
- Oldziej S, Czaplowski C, Liwo A, Chinchio M, Nianis M, Vila JA, Khalili M, Arnavtova YA, Jagielska A, Makowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proc Natl Acad Sci USA* 2005;102:7547–7552.
- Hansmann UH, Okamoto Y. New Monte Carlo algorithms for protein folding. *Curr Opin Struct Biol* 1999;9:177–183.
- Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* 2000;297:233–249.
- Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;5:823–826.
- Hvidsten TR, Kryshstafovich A, Komorowski J, Fidelis K. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics* 2003;19(Suppl 2):II81–II91.
- Bystroff C, Simons KT, Han KF, Baker D. Local sequence-structure correlations in proteins. *Curr Opin Biotechnol* 1996;7:417–421.
- Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* 2000;301:173–190.
- Jones DT. Predicting novel protein folds by using FRAGFOLD. *Proteins* 2001;Suppl 5:127–132.
- Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA* 2001;98:10125–10130.
- Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA* 2004;101:7594–7599.
- Bonneau R, Strauss C, Rohl C, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* 2002;322:65.
- Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. *Curr Opin Chem Biol* 2001;5:51–56.
- Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
- Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000;29:291–325.
- Godzik A. Fold recognition methods. *Methods Biochem Anal* 2003;44:525–546.
- David R, Korenberg MJ, Hunter IW. 3D–1D threading methods for protein fold recognition. *Pharmacogenomics* 2000;1:445–455.
- Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins* 2001;42:319–331.
- Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* 2004;56:502–518.
- Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* 2004;55:1005–1013.
- Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. *Proteins* 2003;53(Suppl):6395–6409.
- Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 2005;15:285–289.
- Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. *Curr Opin Struct Biol* 2002;12:176–181.
- Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. Ab initio protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. *Proc Natl Acad Sci USA* 2002;99:5993–5998.
- Gribkov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987;84:4355–4358.
- Wallace IM, Blackshields G, Higgins DG. Multiple sequence alignments. *Curr Opin Struct Biol* 2005;15:261–266.
- Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. *Protein Sci* 2000;9:1487–1496.
- Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. *Protein Sci* 2004;13:1071–1087.
- Wang G, Dunbrack RL Jr. Scoring profile-to-profile sequence alignments. *Protein Sci* 2004;13:1612–1626.
- Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. *Nucleic Acids Res* 2005;33:1874–1891.
- Tomii K, Akiyama Y. FORTE: a profile–profile comparison tool for protein fold recognition. *Bioinformatics* 2004;20:594–595.
- Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* 2000;9:232–241.
- Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 2000;299:499–520.
- Pawlowski K, Rychlewski L, Zhang B, Godzik A. Fold predictions for bacterial genomes. *J Struct Biol* 2001;134:219–231.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
- Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
- Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. *Methods Enzymol* 1983;91:524–545.
- Prlc A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng* 2000;13:545–550.
- Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci* 1992;1:216–226.
- Qian B, Goldstein RA. Optimization of a new score function for the generation of accurate alignments. *Proteins* 2002;48:605–610.
- Hourai Y, Akutsu T, Akiyama Y. Optimizing substitution matrices by separating score distributions. *Bioinformatics* 2004;20:863–873.
- Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 1990;213:859–883.
- Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasicheical approximation correct? *Protein Sci* 1997;6:676–688.
- Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res* 2000;28:374.
- Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res* 2003;31:492–493.
- Mattice WL, Suter UW. Conformational theory of large molecules. New York: John Wiley & Sons, Inc., 1994.
- Kolaskar AS, Kulkarni-Kale U. Sequence alignment approach to pick up conformationally similar protein fragments. *J Mol Biol* 1992;223:1053–1061.
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C, 2nd ed. New York: Cambridge University Press, 1988.
- Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;85:2444–2448.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
- Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. *J Mol Biol* 2001;307:721–735.
- Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
- Johnson MS, Overington JP. A structural basis for sequence

- comparisons. An evaluation of scoring methodologies. *J Mol Biol* 1993;233:716–738.
54. Koshi JM, Goldstein RA. Context-dependent optimal substitution matrices. *Protein Eng* 1995;8:641–645.
  55. Miyazawa S, Jernigan RL. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Eng* 1993;6:267–278.
  56. Qu CX, Lai LH, Xu XJ, Tang YQ. Phyletic relationships of protein structures based on spatial preference of residues. *J Mol Evol* 1993;36:67–78.
  57. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. *Science* 1992;256:1443–1445.
  58. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol* 2000;295:613–625.
  59. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
  60. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
  61. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998;11:739–747.
  62. Thompson JD, Plewniak F, Poch O. BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 1999;15:87–88.
  63. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.
  64. Stebbings LA, Mizuguchi K. HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res* 2004;32(Database issue):D203–D207.
  65. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;48:443–453.
  66. Saqi MA, Sternberg MJ. A simple method to generate non-trivial alternate alignments of protein sequences. *J Mol Biol* 1991;219:727–732.
  67. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 2001;44:133–149.
  68. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. *J Mol Biol* 2001;307:721–735.
  69. Friedberg I, Jaroszewski L, Ye Y, Godzik A. The interplay of fold recognition and experimental structure determination in structural genomics. *Curr Opin Struct Biol* 2004;14:307–312.
  70. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science* 2005;309:1868–1871.
  71. Felsenstein J. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* 2005;5:164–166.